12

Correlation

Syllabus reference

Data analysis 7

- Correlation

Correlation is a term used in mathematics to describe certain relationships between variables. This chapter focuses on examining and describing the relationship that may exist between two variables and, if a relationship exists, making predictions about one variable if we know the other.

ARE YOU READY?

Try the questions below. If you have difficulty with any of them, extra help can be obtained by completing the matching SkillSHEET. Either click on the SkillSHEET icon next to the question on the *Maths Quest HSC Course* eBookPLUS or ask your teacher for a copy.

eBookplus

Digital doc
SkillSHEET 12.1
doc-1421
Finding the median

Finding the median

1 Find the median of:

a 3, 5, 6, 3, 4, 2, 5, 2, 7

b 12, 15, 10, 11, 15, 15, 16, 11, 19, 16.

eBookplus

Digital doc
SkillSHEET 12.2
doc-1422
Using the regression equation to make predictions

Using the regression equation to make predictions

2 For the equation $y = 5x - 2$ find:

a y if $x = 40$

b x if $y = 258$.

eBookplus

Digital doc
SkillSHEET 12.3
doc-1423
Finding the gradient I

Finding the gradient I

3 Find the gradient of the line joining the points:

a (1, 3) and (4, 12)

b (–2, –4) and (6, –2).

eBookplus

Digital doc
SkillSHEET 12.4
doc-1424
Finding the gradient II

Finding the gradient II

4 Calculate the gradient of the following lines, and state whether the gradient is positive or negative.

a Vertical rise = 12, horizontal run = 2

b Vertical rise = –6, horizontal run = 4

12A Scatterplots

The manager of a small ski resort has a problem. He wants to be able to predict the number of skiers using his resort each weekend in advance, so that he can organise additional resort staffing and catering if needed. He knows that good deep snow will attract skiers in big numbers but scant covering is unlikely to attract a crowd. To investigate the situation further, he collects the following data over twelve consecutive weekends at his resort.

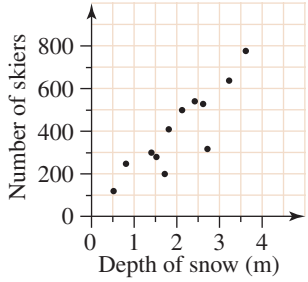
Depth of snow (m)	Number of skiers
0.5	120
0.8	250
2.1	500
3.6	780
1.4	300
1.5	280
1.8	410
2.7	320
3.2	640
2.4	540
2.6	530
1.7	200

As there are two types of data in this example, they are called bivariate data. For each item (weekend), two variables are considered (depth of snow and number of skiers). When analysing bivariate data, we are interested in examining the relationship between the two variables. In the case of the ski resort data we might be interested in answering the following questions.

- Are visitor numbers related to depth of snow?
- If there is a relationship between visitor numbers and depth of snow, is it always true? or is it just a guide? In other words, how strong is the relationship?
- How much confidence could be placed in the prediction?

To help answer these questions, the data can be arranged on a **scatterplot**.

Each of the data points is represented by a single visible point on the graph.



When drawing a scatterplot, it is important to choose the correct variable to assign to each of the axes. The convention is to place the independent variable on the x-axis and the dependent variable on the y-axis. The independent variable in an experiment or investigation is the variable that is deliberately controlled or adjusted by the investigator. The dependent variable is the variable that responds to changes in the independent variable.

Neither of the variables involved in the ski resort data was controlled directly by the investigator, but 'Number of skiers' would be considered the dependent variable because it is likely to change depending on depth of snow. (The snow depth does not depend on numbers of skiers.) As 'Number of skiers' is the dependent variable, we graph it on the y-axis and the 'Depth of snow' on the x-axis.

Notice how the scatterplot for the ski resort data shows a general upward trend. It is not a perfectly straight line, but it is still clear that a general trend or relationship has formed: as the depth of snow increases, so too does the number of skiers.

WORKED EXAMPLE 1

The table below shows the height and mass of ten Year 12 students.

Height (cm)	120	124	130	135	142	148	160	164	170	175
Mass (kg)	45	50	54	59	60	65	70	78	75	80

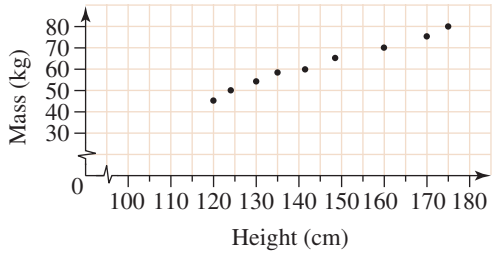
Display the data on a scatterplot.

THINK

WRITE

Method 1: Technology-free

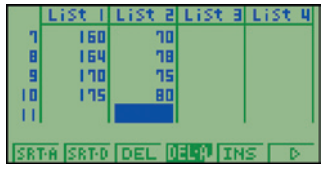
1. Show the height on the horizontal axis and the mass on the vertical axis.
2. Plot the point given by each pair.



Method 2: Technology-enabled

1. From the **MENU** select **STAT**.

2. Delete any existing data, and store the data for height in **List 1** and mass in **List 2**.



3. Press F1 **(GRPH)** (you may have to press F6 for more options first); then press F6 **(SET)**. Set the graph type to **Scatter** by arrowing down to graph type and pressing F1 **(Scat)** (again you may have to press F6 for more options first). Ensure that **XList** is **List 1**, **YList** is **List 2** and **Frequency** is **1** as shown at right.

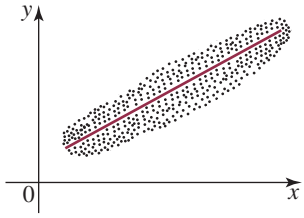
4. Press EXIT to return to the previous screen, and then press F1 **(GPH1)**. The scatterplot will then be drawn.

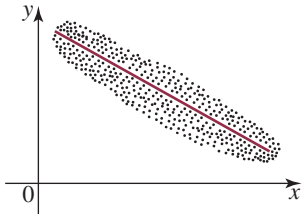
Note that the graphics calculator sets the values on the x- and y-axes automatically. You can press SHIFT F3 **(V-Window)** to set the scale as you see fit.

Once the scatterplot has been drawn, we can determine if any pattern is evident. Worked example 1 shows how, as a general rule, as height increases so does mass.

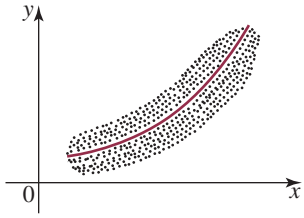
We can also look to see if the pattern is linear. In worked example 1, although the points are not in a perfect straight line, they approximate a straight line. The figures below show examples of linear and non-linear relationships.

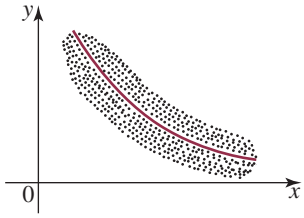
Linear relationships

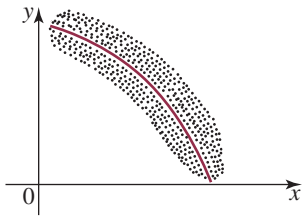


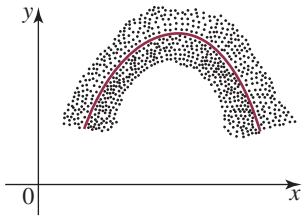


Non-linear relationships

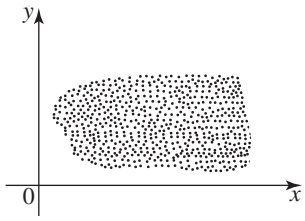








In other cases it may be that there is no relationship at all between the two variables. Such a scatterplot would look like the one shown on the below.



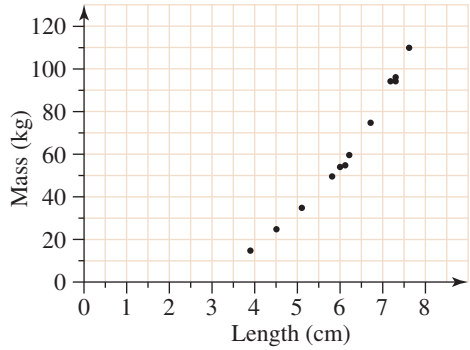
WORKED EXAMPLE 2

The table below shows the length and mass of a dozen eggs.

Length (cm)	6.2	3.9	4.5	5.8	7.2	7.6	6.1	6.7	7.3	5.1	6.0	7.3
Mass (g)	60	15	25	50	95	110	55	75	95	35	54	96

a **Display this information in a scatterplot.**

b **Determine if there is any relationship between the length and mass of the eggs and state if the relationship is linear.**

THINK	WRITE
a 1 Display length on the x-axis and mass on the y-axis.	**a** 
2 Plot the point given by each pair.	
b 1 Study the scatterplot to see if mass increases as length increases.	**b** As length increases, so does the mass of the egg.
2 Study the scatterplot to see if the points seem to approximate a straight line.	The points do not approximate a straight line, and so the relationship is not linear.

REMEMBER

1. A scatterplot is a graph that is used to compare two variables.
2. One variable (the independent variable) is on the horizontal axis, and the other variable (the dependent variable) is on the vertical axis.
3. Points are plotted by the pair formed by each variable.
4. A relationship between the variables exists if one increases as the other increases or if one decreases as the other increases.
5. If the points on the scatterplot seem to approximate a straight line, the relationship can be said to be linear.

EXERCISE

12A Scatterplots

eBookplus
Digital doc
EXCEL Spreadsheet
doc-1417
Scatterplot

1 WE1 The table below shows the marks obtained by a group of ten students in History and Geography. Display this information on a scatterplot.

History	36	65	82	72	58	39	58	74	82	66
Geography	45	78	66	72	50	51	61	70	60	88

eBookplus
Digital doc
EXCEL Spreadsheet
doc-1418
Two variable statistics

2 The table below shows the maximum temperature each day, together with the number of people who attend the cinema that day. Display the information on a scatterplot.

Temperature (°C)	25	33	30	22	15	18	27	22	28	20
No. at cinema	256	184	190	312	458	401	200	357	312	423

3 The table below shows the wages, W, of 20 people and the amount of money they spend each week on entertainment, E. Display this information in a scatterplot.

Wages ($)	370	380	500	510	395	430	535	490	495	550
Amount spent on entertainment ($)	55	85	150	75	145	100	130	115	70	150
Wages ($)	810	460	475	520	530	475	610	780	350	460
Amount spent on entertainment ($)	220	50	100	150	140	160	90	130	40	50

4 WE2 The table below shows the marks obtained by nine students in English and History.

English	55	20	27	33	73	18	37	51	79
History	72	37	53	74	73	44	59	55	84

a Display the information on a scatterplot.
b Is there any relationship between the mark obtained in English and in History? If there does appear to be a relationship, is the relationship linear?

5 The table below shows the daily temperature and the number of hot pies sold at the school canteen.

Temperature (°C)	24	32	28	23	16	14	26	20	29	21
No. of pies sold	56	20	24	60	84	120	70	95	36	63

a Display the information on a scatterplot.
b Determine if there appears to be any relationship between the two variables and if the relationship appears to be linear.

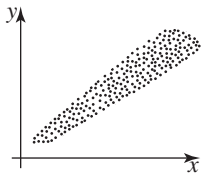
6 Container ships arriving on a wharf are unloaded by work teams. The table below shows the number of people in the work team and the time taken to unload the container ship.

No. in work team	**Hours taken**
15	20
18	16
12	25
19	15
22	14
21	13
17	18
16	20
18	17
20	14

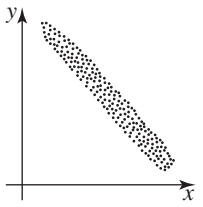
a Display the information on a scatterplot.
b Determine if there appears to be a relationship between the number of people in the work team and the time taken to unload the container ship. If there is a relationship, does the relationship appear to be linear?

7 MC Which of the following scatterplots does not display a linear relationship?

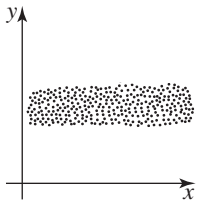
A



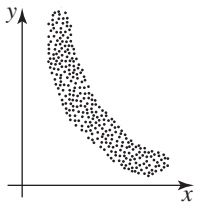
B



C

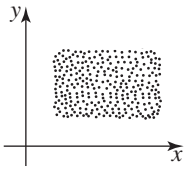


D

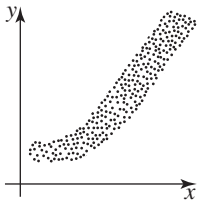


8 MC In which of the following is no relationship evident between the variables?

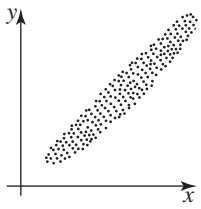
A



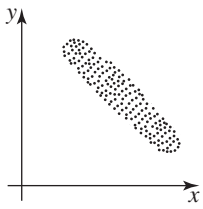
B



C

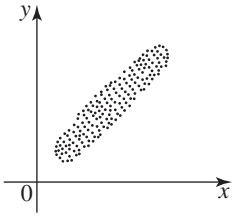


D

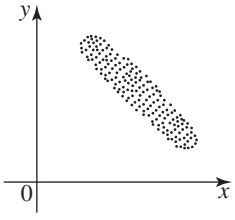


9 Give an example of a situation where the scatterplot may look like the ones below.

a



b



Further development

10 If a relationship appears to exist and if one quantity increases and the other also increases, then the relationship is said to be a positive one. If as one quantity increases the other decreases, then the relationship is said to be negative.

For each of the following state whether you would expect a relationship to exist and if so whether it would be positive or negative.

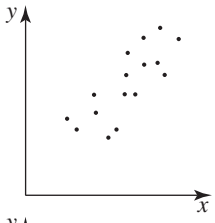
a Time spent studying and the marks achieved
b The number of hours spent training for a cricket team and the number of runs scored
c Age of a person and income level
d Amount spent each week on groceries and the number of hours television watched
e The amount spent on petrol each week and the distance driven

11 For each of the scatterplots drawn below state:

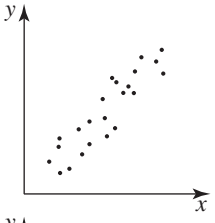
i if a linear relationship exists

ii if a relationship exists, whether that relationship is positive or negative.

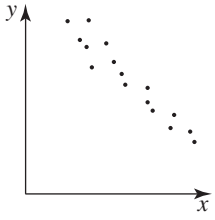
a



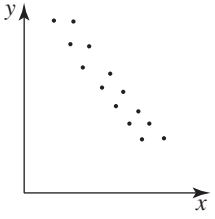
b



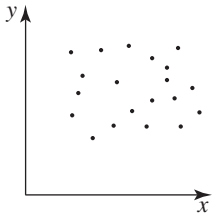
c



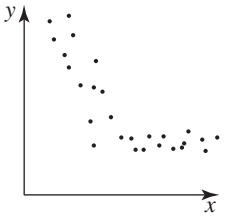
d



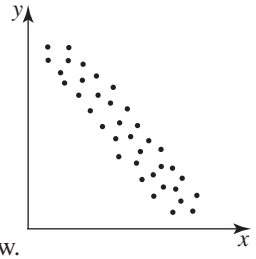
e



f



12 MC From the scatterplot below it appears that:



A there is a negative non-linear relationship.

B there is a positive non-linear relationship.

C there is a negative linear relationship.

D there is a positive linear relationship.

13 The population of each suburb is counted along with the number of supermarkets that are in the area. The results are shown in the table below.

Population (× 1000)	55	65	65	70	75	80	85	85	90	90	95
No. of supermarkets	4	4	6	5	6	8	6	7	8	9	8

Construct a scatterplot of the data and use it to state if there appears to be a linear relationship and whether that relationship is positive or negative.

14 The table below contains data for the time taken to do a concreting job and the cost of the job.

Time taken (hours)	**Cost of job ($)**
5	1500
7	1500
5	2250
8	1800
10	3000
13	2500
15	4200
20	4800
18	4200
25	6000
33	4500

a Display the results in a scatterplot.

b Determine if a linear relationship exists and if so whether it is positive or negative.

15 The table below shows the number of days before a theatre performance booking is made and the best available row number.

Number of days prior	1	2	3	3	4	7	9	10	10	13	16	17	18	20	21	24	25	25	26	27
Best available row number	15	15	15	14	14	13	13	12	10	11	10	8	5	4	3	2	2	1	1	1

a Display the results in a scatterplot.
b Determine if a relationship between the variables exists and if so state whether it is linear and positive or negative.

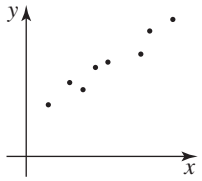
eBookplus
Investigation
Collecting bivariate data
doc-1419

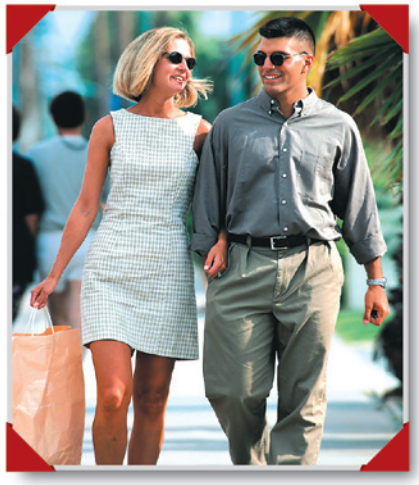
Regression lines

The process of 'fitting' straight lines to bivariate data enables us to analyse relationships between the data and possibly make predictions based on the given data set.

12B Fitting a straight line by eye

Consider the set of *bivariate* data points shown at right. In this case the x-values could be heights of married women, while y-values could be the heights of their husbands. We wish to determine a linear relationship between these two random variables.





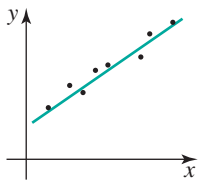
Of course, there is no single straight line that would go through all the points, so we can only *estimate* such a line.

Furthermore, the more closely the points appear to be on or near a straight line, the more confident we are that such a linear relationship may exist and the more accurate our fitted line should be.

Consider the estimate, drawn by eye in the figure below. It is clear that most of the points are on or very close to this straight line. This line was easily drawn since the points are very much part of an apparent linear relationship.

However, note that some points are below the line and some are above it. Furthermore, if x is the height of wives and y is the height of husbands, it seems that husbands are generally *taller* than their wives.

Regression analysis is concerned with finding these straight lines using various methods so that the number of points above and below the lines are balanced.

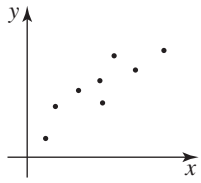


Method of fitting lines by eye

There should be an *equal number of points* above and below the line. For example, if there are 12 points in the data set, 6 should be above the line and 6 below it. This may appear logical or even obvious, but fitting by eye involves a considerable margin of error.

WORKED EXAMPLE 3

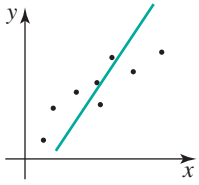
Fit a straight line to the data in the figure using the equal-number-of-points method.

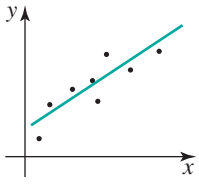


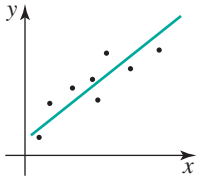
THINK

1. Note that the number of points (n) is 8.
2. Fit a line where 4 points are above the line. Using a clear plastic ruler, try to fit the best line.
3. The first attempt has only 3 points below the line where there should be 4. Make refinements.
4. The second attempt is an improvement, but the line is too close to the points above it. Improve the position of the line until a better balance between upper and lower points is achieved.

DRAW







WORKED EXAMPLE 4

eBook*plus*
Tutorial
int-2441
Worked example 4

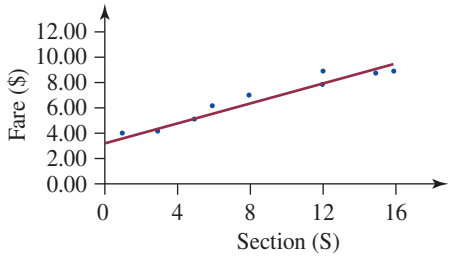
A bus company charges its fares according to the number of sections travelled. The fare for the same number of sections may be different for different routes. The table below shows a variety of fares for a number of sections on different routes.

a **Represent the data in a scatterplot, and draw a line of best fit by eye.**

b **Use your line to find an equation to relate the fare (F) to the number of sections travelled (s).**

c **Explain the meaning of the vertical intercept and gradient of the line in this context.**

Sections	Fare	Sections	Fare
3	$4.20	5	$5.20
1	$4.00	1	$4.00
15	$4.50	16	$9.00
12	$8.00	8	$7.00
12	$9.00	6	$3.00

THINK	WRITE
a 1 Draw the axis with the number of sections on the horizontal axis and the fare on the vertical axis.	**a** 
2 Plot the data and draw your line on the scatterplot.	
b 1 Find the y-intercept.	**b** $b = 3.6$
2 Choose two points on the line and use the gradient formula. It is easiest to make the y-intercept one of your points.	Take (0, 3.6) and (12, 8) $m = \frac{\text{vertical change in position}}{\text{horizontal change in position}}$ $= \frac{5.4}{12}$ $= 0.45$
3 Use the gradient intercept form of a straight line to write your equation.	$F = 0.45s + 3.6$
c 1 The y-intercept is the fare when $s = 0$.	**c** The vertical intercept is the minimum cost of a bus fare.
2 The gradient is the increase in F for every extra section.	The gradient is the cost per section.

REMEMBER

To fit a straight line by eye, when using bivariate data, make sure there are an equal number of points above and below the fitted line.

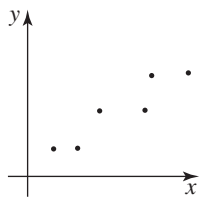
EXERCISE

12B Fitting a straight line by eye

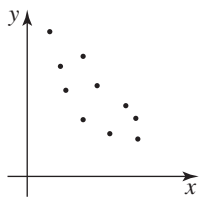
The questions below represent data collected by groups of students conducting different environmental projects. The students have to fit a straight line to their data sets.
Note: For many of these questions your answers may differ somewhat from those in the back of the book. The answers are provided as a guide but there are likely to be individual differences when fitting straight lines by eye.

1 WE3 Fit a straight line to the data in the scatterplots using the equal-number-of-points method.

a

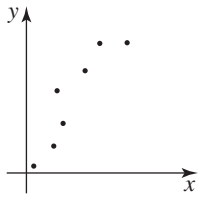


b

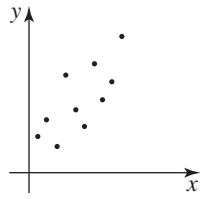


eBookplus
Digital doc
WorkSHEET 12.1
doc-1420

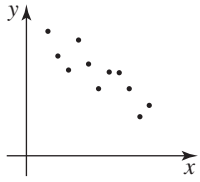
c



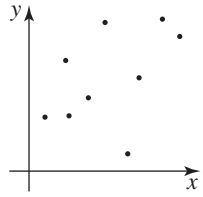
d



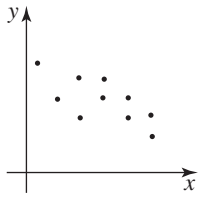
e



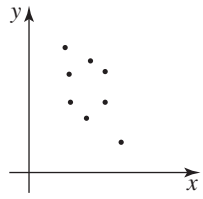
f



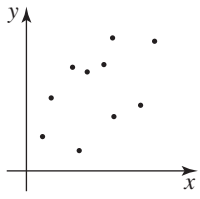
g



h



i

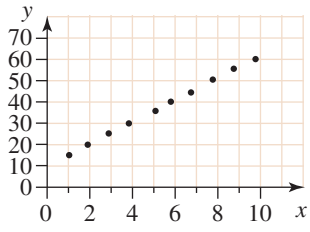


2 For each of the following draw a line of best fit and for each find:

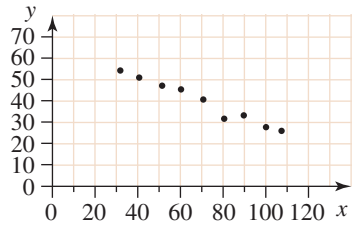
i the gradient

ii the vertical intercept.

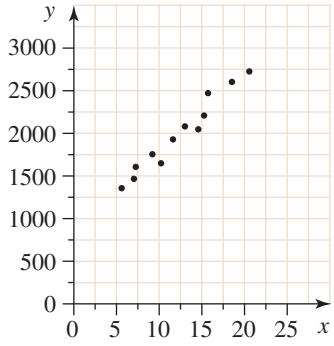
a



b



c



3 WE4 The table below shows the length of an elastic when stretched by a force expressed in newtons.

Force	0	1	2	3	4	5	6	7	8	9
Length	440	450	462	470	484	492	500	508	518	528

a Represent the data in a scatterplot, and draw a line of best fit by eye.
b Use your line to find an equation to relate the length (L) to the force (f).
c Explain the meaning of the vertical intercept and gradient of the line in this context.

Further development

4 A solid substance has its mass measured at various temperatures. The results are shown in the table below.

Temperature (°C)	–20	–15	–10	0	5	10	15	20	25	30
Mass (kg)	2.4	3.8	4.8	6.2	7.2	8.2	9.6	10.6	12.2	13.4

a Represent the data in a scatterplot, and draw a line of best fit by eye.
b Use your line to find an equation to relate the mass (M) to the temperature (t).
c Explain the meaning of the vertical intercept and gradient of the line in this context.

12C Fitting a straight line — the 3-median method

Fitting lines by eye is useful, but it is not the most accurate of methods. Greater accuracy is achieved through closer analysis of the data. Upon closer analysis it is possible to find the equation of a **line of best fit** of the form $y = mx + c$, where m is the gradient and c is the y-intercept. Several mathematical methods provide a line with a more accurate fit.

One of these methods is called the *3-median method* and involves the division of the data set into 3 groups, and the use of the 3 medians in these groups to determine a line of best fit. It is used when data show a linear relationship. It can even be used when the data contain outliers. The 3-median method is best described as a step-by-step method.

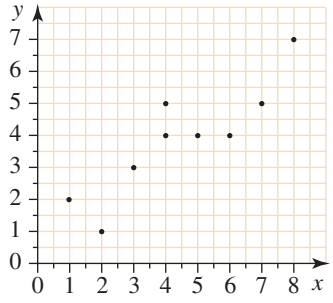


Figure 1

Step 1. Plot the points on a scatter diagram. This is shown in figure 1.

Step 2. Divide the points into 3 groups using vertical divisions (see figure 2). The number of points in a data set will not always be exactly divisible by 3. Thus, there will be three alternatives, as follows.
(a) If the number of points is divisible by 3, divide them into 3 equal groups, for example, 3, 3, 3 or 7, 7, 7.
(b) If there is 1 extra point, put the extra point in the middle group, for example, 3, 4, 3 or 7, 8, 7.
(c) If there are 2 extra points, put 1 extra point in each of the outer groups, for example, 4, 3, 4 or 8, 7, 8.

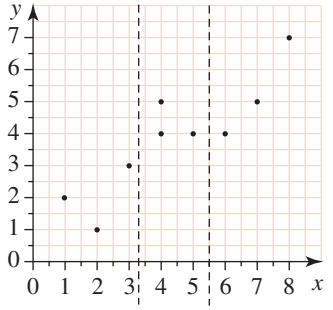


Figure 2

Step 3. Find the median point of each of the 3 groups and mark each median on the scatterplot (see figure 3). Recall that the median is the middle value. So the

median point of each group has an x-coordinate that is the median of the x-values in the group and a y-coordinate that is the median of the y-values in the group.

(a) The left group is the *lower* group and its median is denoted by (x_L, y_L).

(b) The median of the *middle* group is denoted by (x_M, y_M).

(c) The right group is the *upper* group and its median is denoted by (x_U, y_U).

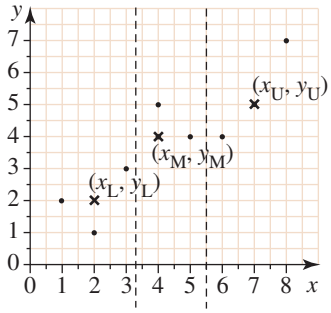


Figure 3

Note: Although the x-values are already in ascending order on the scatterplot, the y-values within each group may need re-ordering before you can find the median.

Step 4. Draw in the line of best fit. Place your ruler so that it passes through the lower and upper medians. Move the ruler a third of the way toward the middle group median *while maintaining the slope*. Hold the ruler there and draw the line.

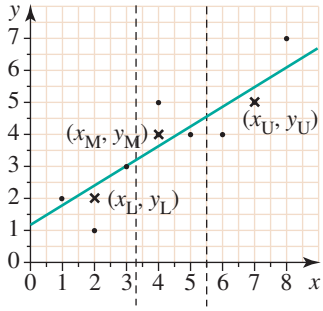


Figure 4

Step 5. Find the equation of the **3-median regression line** (general form $y = mx + b$). Draw on your knowledge of finding equations of lines to find the equation of the line drawn on the scatterplot. If the scale on the axes begins at zero, you can read off the y-intercept of the line and calculate the gradient of the line. This will enable you to find the equation of the line.

The equation of a straight line can be found using $y = mx + b$, where m is the gradient and b is the y-intercept. The gradient of the regression line is best found with a ruler and using the formula:

$$m = \frac{\text{vertical change in position}}{\text{horizontal change in position}}$$

WORKED EXAMPLE 5

eBook*plus*
Tutorial
int-2442
Worked example 5

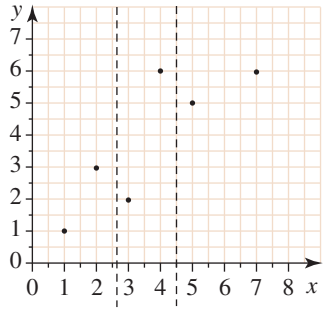
Find the equation of the regression line for the data in the table below using the 3-median method.

x	1	2	3	4	5	7
y	1	3	2	6	5	6

THINK

1 Plot the points on a scatterplot, and divide the data into 3 groups. Note there are 6 points, so the division will be 2, 2, 2.

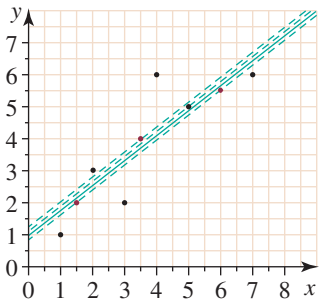
WRITE



2 Find the median point of each group. Since each group has only 2 points, medians are found by averaging them.

$(x_L, y_L) = (1.5, 2)$
$(x_M, y_M) = (3.5, 4)$
$(x_U, y_U) = (6, 5.5)$

3 Mark in the medians, and place a ruler on the outer 2 medians. Maintaining the same slope on the ruler, move it one-third of the way towards the middle median. Draw the line.



4 Read off the y-intercept from the graph.

y-intercept = 1

5 Use (x_L, y_L) and (x_U, y_U) to calculate the gradient.

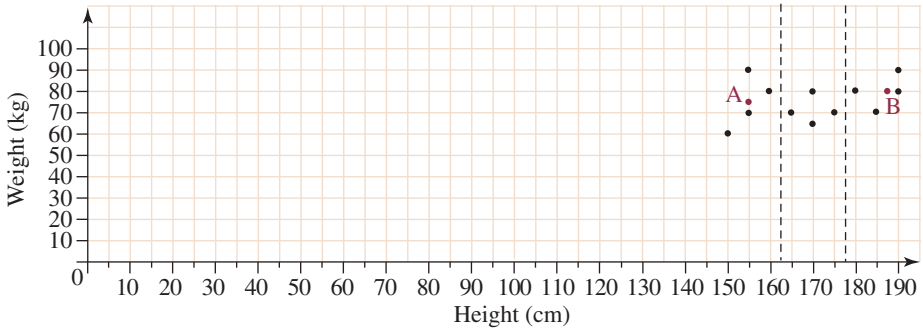
$$\text{Gradient } (m) = \frac{5.5-2}{6-1.5}$$
$$= \frac{3.5}{4.5}$$
$$= \frac{7}{9}$$

6 Write the equation of the 3-median regression line.

$$y = \frac{7}{9}x + 9 \quad \text{or}$$
$$9y = 7x + 81$$

WORKED EXAMPLE 6

The scatterplot below shows a comparison between the heights and weights of 12 boys. The median points A and B in the first and last sections have been found for you.



a Find the coordinates of median point C, and hence find the median regression line.
b Find the gradient and y-intercept of the regression line, and hence find the equation of the regression line.

THINK	WRITE
Method 1: Technology-free	
a 1 Find the coordinates of point C by finding the median of the x-values and finding the median of the y-values.	**a** x-values are: 165, 170, 170 and 175 Median x-value $= \frac{170+170}{2} = 170$

2 Mark point C on the diagram.

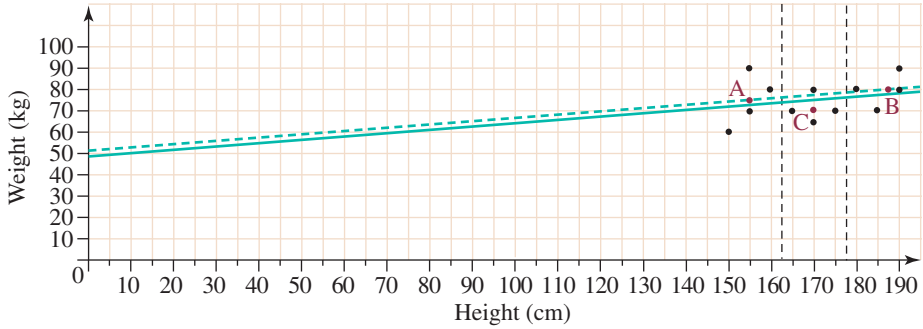
3 Rule a line through points A and B.

4 Move the line AB one-third of the way towards C, keeping the new line parallel to AB.

y-values are: 65, 70, 70 and 80

Median *y*-value $= \frac{70+70}{2} = 70$

The coordinates of C are (170, 70).



b 1 Calculate the gradient, *m*, by finding the rise and run between two points on the line.

b $m = \frac{\text{rise}}{\text{run}}$

$= \frac{78-70}{190-140}$

$= 0.16$

2 Read the value from the graph to state the *y*-intercept, *b*.

$b = 49$

3 Substitute *m* and *b* into the formula $y = mx + b$ to find the equation of the regression line.

The equation is of the form $y = mx + b$, where *x* represents height in cm and *y* represents weight in kg.

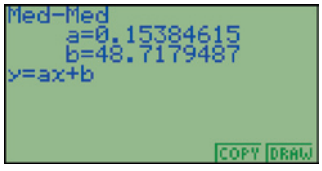
$y = 0.16x + 49$

Method 2: Technology-enabled

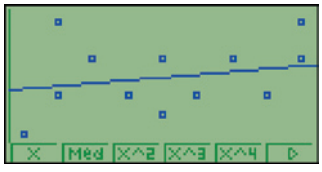
1 From the **MENU** select **STAT**.

2 Enter the data into **List 1** and **List 2** and draw the scatterplot as shown in the previous section. Since we are using the calculator it is not necessary to draw the scatterplot from 0 on the axes.

3 Press F2 **(Med)** to find the equation of the median regression line. The value of **a** is the gradient of the line and the value of **b** is the *y*-intercept.



4 If you want to see the regression line drawn on the scatterplot, press F6 **(DRAW)**.



In the previous example we would give the equation $y = 0.15x + 49$, which is slightly different from the example done on paper. Because the method relies on the eye to find two points on the regression line to find the gradient and y-intercept, minor differences are insignificant and quite acceptable.

Once the regression line has been found, we are able to use the equation to make predictions about other pieces of data.

WORKED EXAMPLE 7

A casino records the number of people, N, playing a jackpot game and the prize money, p, for that game and plots the results on a scatterplot. The regression line is found to have the equation $N = 0.07p + 220$.

a Find the number of people playing when the prize money is $2500.

b Find the likely prize on offer when there are 500 people playing.

THINK		WRITE
a 1	Write the equation of the regression line.	a $N = 0.07p + 220$
2	Substitute 2500 for p.	$N = 0.07 \times 2500 + 220$
3	Calculate N.	$= 395$
4	Give a written answer.	There would be approximately 395 people playing.
b 1	Write the equation of the regression line.	b $N = 0.07p + 220$
2	Substitute 500 for N.	$500 = 0.07p + 220$
3	Solve the equation.	$280 = 0.07p$ $p = 4000$
4	Give a written answer.	The prize would be approximately $4000.

REMEMBER

1. The median regression line is the line of best fit that is drawn on a scatterplot.
2. The median regression line can be drawn using the method of three medians.
3. To find the median regression line:
   (a) divide the points into three approximately equal sections. If the number of points is not divisible by three, make sure there is the same number of points in the first and last sections.
   (b) mark median points in the first and last sections by finding the median of the x-values and finding the median of the y-values for each section. Label these points A and B.
   (c) find the median point in the middle section and label this point C.
   (d) draw the line AB and then move the line one-third of the way towards C, keeping the line parallel to AB.
4. The equation of the regression line can be found by measuring the gradient and the y-intercept of the regression line and using the formula $y = mx + b$. Sometimes the gradient of the median regression line will be negative.
5. Once the equation of the regression line has been found, it can then be used to make predictions about the variables.

EXERCISE

12C Fitting a straight line — the 3-median method

eBookplus
Digital doc
SkillSHEET 12.1
doc-1421
Finding the median

1 WE5 The table below shows the marks achieved by a class of students in English and Maths.

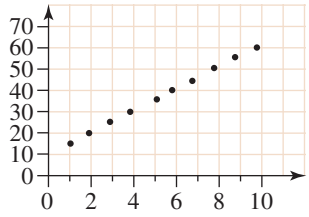
English	64	75	81	63	32	56	47	59	73	64
Maths	76	62	89	56	49	57	53	72	80	50

Show these data on a scatterplot, and on the graph show the regression line using the 3-median method.

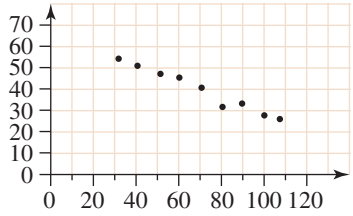
eBookplus
Digital doc
SkillSHEET 12.2
doc-1422
Using the regression equation to make predictions

2 Position the median regression line, using the 3-median method, through each of the following graphs, and find the equation of each.

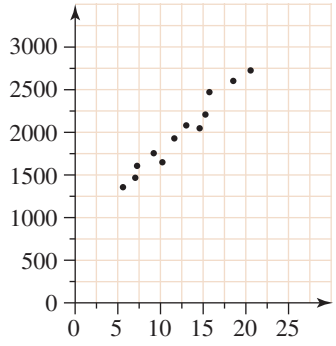
a



b



c



eBookplus
Digital doc
SkillSHEET 12.3
doc-1423
Finding the gradient I

eBookplus
Digital doc
SkillSHEET 12.4
doc-1424
Finding the gradient II

3 WE6 In an experiment, a student measures the length of a spring when different masses are attached to it. Her results are shown below.

Mass (g)	0	100	200	300	400	500	600	700	800	900
Length of spring (mm)	220	225	231	235	242	246	250	254	259	264

eBookplus
Digital doc
EXCEL Spreadsheet
doc-1425
3-median regression

a Draw a scatterplot of the data, and on it draw the median line of regression, using the 3-median method.

b Find the gradient and y-intercept of the regression line, and hence find the equation of the regression line.

eBookplus
Digital doc
EXCEL Spreadsheet
doc-1426
Making predictions

4 A scientist who measures the volume of a gas at different temperatures provides the table of values at right.

a Draw a scatterplot of the data and on it draw the line of regression using the 3-median method.

b Give the equation of the line of best fit. Write your equation in terms of the variables: volume of gas, V, and its temperature, T.

Temperature (°C)	Volume (L)
–40	1.2
–30	1.9
–20	2.4
0	3.1
10	3.6
20	4.1
30	4.8
40	5.3
50	6.1
60	6.7

5 A sports scientist is interested in the importance of muscle bulk to strength. He measures the biceps circumference of ten people and tests their strength by asking them to complete a lift test. His results are given in the following table.

Circumference of biceps (cm)	Lift test (kg)
25	50
25	52
27	58
28	51
30	60
30	62
31	53
33	62
34	61
36	66

a Draw a scatterplot of the data and draw the median line of regression using the 3-median method.
b Find a rule for determining the ability of a person to complete a lift test, S, from the circumference of their biceps, B.

6 WE7 A taxi company adjusts its meters so that the fare is charged according to the following equation: $F = 1.2d + 3$, where F is the fare, in dollars, and d is the distance travelled, in km.
a Find the fare charged for a distance of 12 km.
b Find the fare charged for a distance of 4.5 km.
c Find the distance that could be covered on a fare of $27.
d Find the distance that could be covered on a fare of $13.20.

7 Detectives can use the equation $H = 6.1f - 5$ to estimate the height of a burglar who leaves footprints behind. (H is the height of the burglar, in cm, and f is the length of the footprint.)
a Find the height of a burglar whose footprint is 27 cm in length.
b Find the height of a burglar whose footprint is 30 cm in length.
c Find the footprint length of a burglar of height 185 cm. (Give your answer correct to 2 decimal places.)
d Find the footprint length of a burglar of height 152 cm. (Give your answer correct to 2 decimal places.)

8 A pie seller at a football match finds that the number of pies sold is related to the temperature of the day. The situation could be modelled by the equation $N = 870 - 23t$, where N is the number of pies sold and t is the temperature of the day.
a Find the number of pies sold if the temperature was 5 degrees.
b Find the number of pies sold if the temperature was 25 degrees.
c Find the likely temperature if 400 pies were sold.
d How hot would the day have to be before the pie seller sold no pies at all?

9 The following table shows the average annual costs of running a car. It includes all fixed costs (registration, insurance etc.) as well as running costs (petrol, repairs etc.).

Distance (km)	Annual cost ($)
5000	4000
10 000	6400
15 000	8400
20 000	10 400
25 000	12 400
30 000	14 400

a Draw a scatterplot of the data.
b Using the 3-median method, draw in the line of best fit.
c Find an equation which represents the relationship between the cost of running a vehicle, C, and the distance travelled, d.
d Use your graph and its equation to find:
  i the annual cost of running a car if it is driven 15 000 km
  ii the annual cost of running a car if it is driven 1000 km
  iii the likely number of kilometres driven if the annual costs were $8000
  iv the likely number of kilometres driven if the annual costs were $16 000.

10 A market researcher finds that the number of people who would purchase 'Wise-up' (the thinking man's deodorant) is related to its price. He provides the table of values at right.

Price ($)	Weekly sales (× 1000)
1.40	105
1.60	101
1.80	97
2.00	93
2.20	89
2.40	85
2.60	81
2.80	77
3.00	73
3.20	69
3.40	65

a Draw a scatterplot of the data.
b Draw in the line of best fit.
c Find an equation that represents the relationship between the number of cans of 'Wise-up' sold, N (in thousands), and its price, p.
d Use the equation to predict the number of cans sold each week if:
  i the price was $3.10
  ii the price was $4.60.
e At what price should 'Wise-up' be sold if the manufacturers wished to sell 80 000 cans?
f Given that the manufacturers of 'Wise-up' can produce only 100 000 cans each week, at what price should it be sold to maximise production?

11 The following table gives the adult return air fares between some Australian cities.

City	Distance (km)	Price ($)
Melbourne–Sydney	713	580
Perth–Melbourne	2728	1490
Adelaide–Sydney	1172	790
Brisbane–Melbourne	1370	890
Hobart–Melbourne	559	520
Hobart–Adelaide	1144	820
Adelaide–Melbourne	669	570

a Draw a scatterplot of the data and on it draw the median regression line using the line of best fit.
b Find an equation that represents the relationship between the air fare, A, and the distance travelled, d.
c Use the equation to predict the likely air fare (to the nearest dollar) from:
  i Sydney to the Gold Coast (671 km)
  ii Perth to Adelaide (2125 km)
  iii Hobart to Sydney (1024 km)
  iv Perth to Sydney (3295 km).

12 Rock lobsters (crayfish) are sized according to the length of their carapace (main body shell). The table below gives the age and carapace length of 16 male rock lobsters.

Age (years)	Length of carapace (mm)
3	65
2.5	59
4.5	80
4.5	80
3.25	68
7.75	130
8	150
6.5	112
12	200
14	210
4.5	82
3.5	74
2.25	51
1.76	48
10	171
9.5	160

a Display this information on a scatterplot, and on your scatterplot draw the median line of regression using the line of best fit.
b Find the equation of the median regression line.
c Use the equation to find the likely size of a 5-year-old male rock lobster.
d Use the equation to find the likely size of a 16-year-old male rock lobster.
e Rock lobsters reach sexual maturity when their carapace length is approximately 65 mm. Use the equation to find the age of the rock lobster at this stage.
f The fisheries department wants to set minimum size restrictions so that the rock lobsters have three full years from the time of sexual maturity in which to breed before they can be legally caught. What size should govern the taking of a male rock lobster?
Note: Answers for this exercise are approximate and may vary due to the precise location of the line of best fit.

eBookplus
Investigation
Relationship between variables
doc-1427

Further development

13 Copy and complete the table below, which shows the division of data points into three groups for drawing the three-median regression line.

Total number of points	Lower third	Middle third	Upper third
7	2	3	2
8			
9			
10			
11			
12			

(continued)

Total number of points	Lower third	Middle third	Upper third
13			
14			
15			
20			
25			
50			

14 MC When using the three-median method for fitting a regression line, which of the following statements is not correct?

- **A** The gradient of the line can be found using only the two outside medians.
- **B** The vertical intercept is found by moving the line one-third of the way towards the centre median.
- **C** The gradient will remain unchanged when moving the line towards the centre median.
- **D** There must be an equal number of points in each group.

15 The figures below show the number of people (in thousands) that attend on 10 days of the Easter Show.

Day	1	2	3	4	5	6	7	8	9	10
Attendance (× 1000)	75	67	71	63	58	62	54	47	43	39

Find the equation of the three-median regression line.

16 The table below shows the population (P) of nine regional towns together with the percentage of people in those towns who have experience in the farming industry (F).

Population (× 1000)	23	44	77	100	112	156	177	199	213
Farming industry experience (%)	12.3	11.6	9.4	9.6	8.1	8.2	6.2	5.4	4.5

Find the equation of the 3-median regression line.

17 During an experiment the following data was gathered connecting two variables x and y.

x	0	1	2	3	4	5	6	7	8	9	10
y	2	4	5	8	10	15	14	12	18	21	25

x	11	12	13	14	15	16	17	18	19	20	
y	23	27	30	29	31	35	28	38	39	43	

a Display the data as a scatterplot.

b Find the equation of the three-median regression line.

12D Correlation

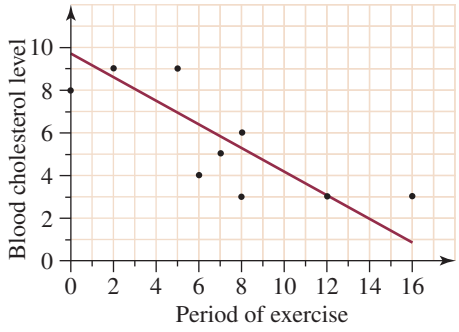
Correlation is a description of the relationship that exists between two variables. When one variable increases with another, it is said that there is a positive correlation between the variables. In such a case, the median regression line will have a positive gradient. Similarly, if one variable decreases while the other increases, the median regression line will have a negative gradient and the correlation is negative.

Consider the following example in which ten Year 11 students were surveyed to find the amount of time that they spend doing exercise each week. This was compared with their blood cholesterol level.

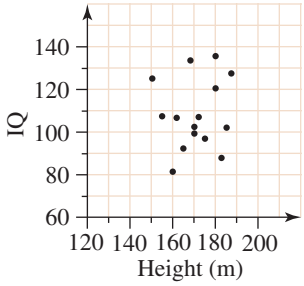
Period of exercise (h)	6	8	12	16	2	0	5	8	7	12
Blood cholesterol level	4	3	3	3	9	8	9	6	5	4

In this example there seems to be a general downward trend, and the median regression line therefore has a negative gradient. As the amount of exercise increases, the level of blood cholesterol decreases.

Notice that in this case the points are not as closely aligned as in the previous examples. We can say that the relationship (or correlation) between the variables is only weak. In general terms, the closer that the points are to forming a straight line, the stronger the relationship is between the variables.

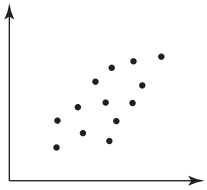


Sometimes we find that there is no relationship between the variables. In the scatterplot below, a researcher was looking for a link between people's heights and their IQs. The points appear to be randomly dispersed across the scatterplot. In cases like this, it can be concluded that there is no clear relationship between the variables.



WORKED EXAMPLE 8

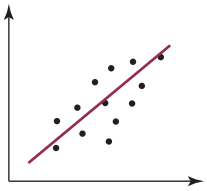
In the figure on the right, describe the correlation as being positive or negative.



THINK

1. Add a median regression line to the scatterplot.

WRITE

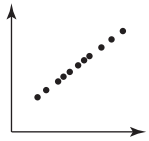
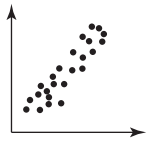
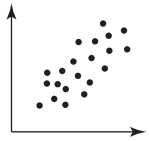
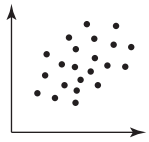
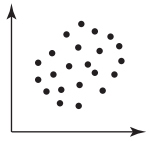
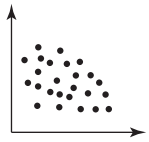
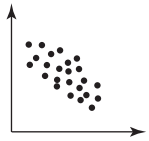
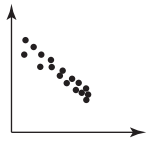
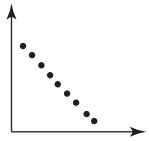


2. The gradient of the regression line is positive.

There is a positive correlation.

3. Therefore the correlation is positive.

The strength of a correlation is based on the **correlation coefficient**. The correlation coefficient is a measure of a correlation.

Correlation coefficient	Description	Scatterplot
1	Perfect positive correlation	
Between 0.75 and 1	Strong positive correlation	
Between 0.5 and 0.75	Moderate positive correlation	
Between 0.25 and 0.5	Weak positive correlation	
Between –0.25 and 0.25	No correlation	
Between –0.5 and –0.25	Weak negative correlation	
Between –0.75 and –0.5	Moderate negative correlation	
Between –1 and –0.75	Strong negative correlation	
–1	Perfect negative correlation	

WORKED EXAMPLE 9

The operators of a casino keep records of the number of people playing a 'Jackpot' type game and compare the numbers playing to the size of the jackpot. The correlation coefficient for this game is calculated to be 0.65. Describe the correlation between the prize and the number of players.

THINK

The correlation coefficient is between 0.5 and 0.75 and so it is a moderate positive correlation.

WRITE

There is a moderate positive correlation between the jackpot and the number of players in the game.

Causality

Causality refers to one variable causing another. For example, there is a high correlation between a person's shoe size and shirt size. However, one does not cause the other. Similarly, there is a high correlation between number of cigarettes smoked and lung cancer but, in this case, smoking causes lung cancer.

Explain whether a positive or negative relationship exists and discuss causality in each of the following.

1. Hours of study and exam marks
2. Hours of exercise and resting pulse rate
3. Weight and shirt size
4. The number of hotels and churches in country towns
5. The number of motels in a town and the number of flights landing at the nearest airport

It is possible to make a qualitative judgement as to the type of correlation that is involved in a relationship by the general appearance of the graph. Care must be taken before making a statement about one variable causing the other.

Just because there is a strong relationship between two variables, it does not mean that one variable causes the other. For example, there is a very strong positive correlation in people between their shoe size and their shirt size, but one does not cause the other.

Similarly, there is a very strong correlation between the amount of study done for an exam and the result achieved on the exam. In this case it can be argued that the study causes the high exam mark. Each case needs to be considered on its merit.

WORKED EXAMPLE 10

A manufacturer who is interested in minimising the cost of training gives 15 of his plant operators different amounts of training and then measures the number of errors made by each of the operators. The results of the experiment are placed on a scatterplot and the correlation between the number of hours of training and the number of errors made is measured to have a correlation coefficient of –0.69.

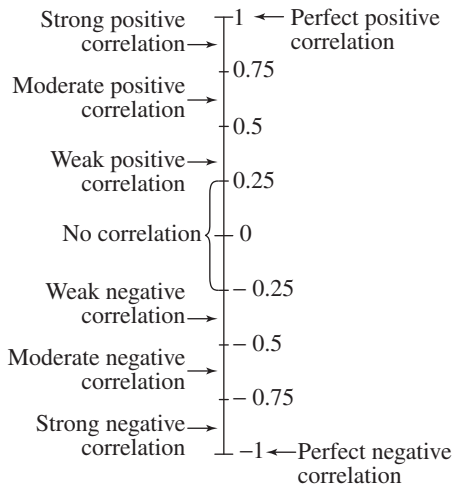
a What can be said of the correlation between training and errors?

b What conclusion could the manufacturer make about causality in this case?

THINK	WRITE
a 1 The correlation coefficient is between –0.75 and –0.5.	**a**
2 A correlation coefficient in this range indicates a moderate negative correlation.	There is a moderate negative correlation between the amount of training and the number of errors made.
b In this case it would seem logical that those that have undertaken more training would make fewer errors.	**b** The manufacturer could reasonably presume that the more training a person is given, the less likely they are to make errors with the machinery.

REMEMBER

1. The pattern of the scatterplot gives an indication of the level of association (correlation) between the variables.
2. When one variable increases with another, there is a positive correlation between them.
3. When one variable decreases while the other increases, there is negative correlation.
4. The extent of the correlation is then measured by the correlation coefficient. The description of the correlation is given in the figure on the right.
5. Strong correlation between two variables does not necessarily mean that one variable causes the other.

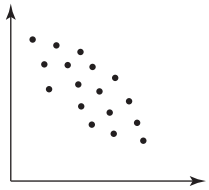


EXERCISE

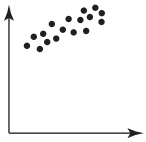
12D Correlation

1 WE8 For each of the following, state whether a positive or negative correlation exists.

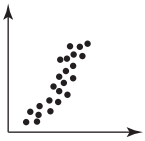
a



b



c



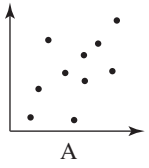
2 A sample of 10 drivers was taken. Each driver was asked their age and the number of speeding offences they had committed in the past five years. The results are in the table below.

Age	22	36	48	40	58	64	23	25	30	45
Speeding offences	4	2	1	1	2	0	3	7	1	0

a Display the information on a scatterplot.
b State if there is a positive or a negative correlation between age and speeding offences.

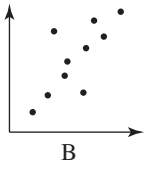
3 Match each of the following scatterplots with the correlation that it shows.

a



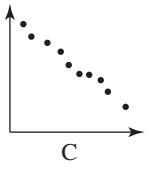
A

b



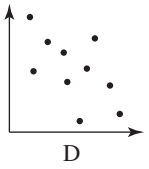
B

c



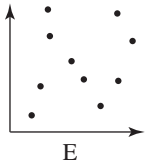
C

d



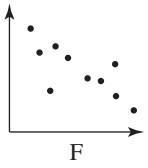
D

e



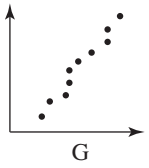
E

f



F

g



G

Strong positive correlation
Weak positive correlation
Weak negative correlation
Strong negative correlation
Moderate positive correlation
No correlation
Moderate negative correlation

4 A pie seller at a football match notices that there seems to be a relationship between the number of pies that he sells and the temperature of the day. He collects the following data.

Daily temperature (°C)	12	22	26	11	8	18	14	16	15	16
Number of pies sold	620	315	295	632	660	487	512	530	546	492

a Draw a scatterplot of the data.
b State the type of correlation that the scatterplot shows and draw a conclusion from the graph.

5 A researcher is investigating the effect of living in airconditioned buildings upon general health. She records the following data.

Hours spent each week in airconditioned buildings	**Number of days sick due to flu and colds**
2	3
13	6
6	2
48	15
40	13
0	8
10	14
0	1
2	16
5	9
18	9
10	6

a Plot the data on a scatterplot.
b State the type of correlation the graph shows and draw a conclusion from it.
c The researcher finishes her experimental report by concluding that airconditioning is the cause of poor health. Is she correct to say this? What other factors could have influenced the relationship shown by the scatterplot?

6 The data below show the population and area of the Australian states and territories.

State	Area ($\times$ 1000 km^2)	Population ($\times$ 1000)
Vic.	228	5092
NSW	802	6828
ACT	2	329
Qld	1727	4053
NT	1346	207
WA	2526	2051
SA	984	1555
Tas.	68	489

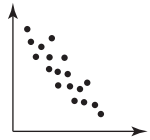
a Plot the data on a scatterplot.
b State the type of correlation the graph shows and draw a conclusion from it.

7 In an experiment, 12 people were administered different doses of a drug. When the drug had taken effect, the time taken for each person to react to a set stimulus was measured. The results are detailed below.

Amount of drug (mg)	Reaction time (s)
0.1	0.030
0.2	0.025
0.3	0.028
0.4	0.036
0.5	0.040
0.6	0.052
0.7	0.046
0.8	0.068
0.9	0.085
1.0	0.092
1.1	0.084
1.2	0.096

a Plot the data on a scatterplot.
b State the type of correlation the graph shows, and draw a conclusion from it.

8 MC What type of correlation is shown by the graph on the right?
A No correlation
B Weak negative correlation
C Moderate negative correlation
D Strong negative correlation



9 MC What type of correlation is shown by the graph on the right?
A No correlation
B Weak positive correlation
C Moderate positive correlation
D Strong positive correlation

10 What type of correlation would be represented by scatterplots that had the following correlation coefficients?

a 1.0	**b** 0.4	**c** 0.8	**d** –0.7
e 0.35	**f** 0.21	**g** –0.75	**h** –0.50
i –0.25	**j** –1.0		

11 WE9 A researcher investigating the proposition that 'tall mothers have tall sons' measures the heights of 12 mothers and the heights of their adult sons. The correlation coefficient is found to be 0.67. Describe the correlation between tall mothers and tall sons.

12 A teacher who is interested in the amount of time students spend doing homework asks 15 students to record the amount of time that they spend on homework and on watching television. The correlation coefficient is found to be –0.45. Interpret the correlation between homework and television watching.

13 A psychologist asked 20 people to rate their 'level of contentment' on a scale of 0 to 10 (10 representing 'perfectly content'). This rating is compared to annual income.

a The correlation coefficient is found to be –0.18. Describe the correlation between income and level of contentment.

b The researcher then intends to write an essay entitled 'Money can't buy happiness'. Do the results confirm this statement?

14 WE10 An experimenter who is investigating the relationship between exercise and obesity measures the weights of 30 boys (of equal height) and also documents the amount of physical exercise that the boys completed each week. The correlation coefficient is found to be –0.47.

a What can be said of the correlation between obesity and exercise?

b What conclusion could be made about causality in this case?

eBookplus
Digital doc
WorkSHEET 12.2
doc-1428

15 MC A researcher is interested in the association between the work rate of production workers and the level of incentive that they are offered under a certain scheme. After drawing a scatterplot, she calculates the correlation between the two variables at 0.82. The researcher can conclude that:

A There is a strong positive correlation between the variables; the greater the incentive, the lower the work rate.

B There is a strong positive correlation between the variables; the greater the incentive, the greater the work rate.

C There is a strong negative correlation between the variables; the greater the incentive, the lower the work rate.

D There is a strong negative correlation; incentives cause an increase in the work rate.

Further development

16 The following proposition is to be investigated: Tall mothers are more likely to have a heavier baby.
The results of a study are shown in the table below.

Height of mother (cm)	Mass of baby (kg)		Height of mother (cm)	Mass of baby (kg)
185	3.76		159	3.20
152	3.24		154	2.96
168	3.36		168	3.56
166	3.32		148	3.04
173	3.58		162	3.68
172	3.44		171	3.60

a Display the information in a scatterplot.
b Describe the correlation between the two variables.

17 The data below shows the relationship between two variables t and y.

t	1	2	3	4	5	6	7	8	9	10	11	12
y	6	9	13	8	9	14	15	17	14	11	15	19

a Display the information in a scatterplot, and fit a regression line by eye.
b Describe the relationship between the variables t and y.

18 The monthly share price of a recently privatised telephone company was recorded as follows.

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.
Price ($)	2.50	2.70	3.00	3.20	3.60	3.70	3.90	4.20

a Graph the data on a scatterplot by letting 1 = Jan., 2 = Feb. etc, and fit a regression line to your data.
b Use your regression line to predict the share price next January.
c Describe the correlation between month and share price.

19 The data below shows monthly sales data for umbrellas.

Month	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Sales	5	10	15	40	70	95	100	90	60	35	20	10

a Describe the correlation between month and sales figures.
b Explain why no regression line can be fitted to this data.

20 The following table shows the quarterly sales figures (in thousands) of a popular software product.

Quarter	2008, Q1	2008, Q2	2008, Q3	2008, Q4	2009, Q1	2009, Q2	2009, Q3	2009, Q4	2010, Q1	2010, Q2	2010, Q3	2010, Q4
Sales	120	135	150	145	140	120	100	110	120	140	190	220

a Put this information on a scatterplot.
b Describe the correlation between quarters and sales.
c Explain why no regression line can be easily fitted to this data.

21 The number of employees at the Comnatpac Bank was recorded over a 10-month period.

Month	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Employees	6100	5700	5400	5200	4800	4400	4200	4000	3700	3300

a Put this data onto a scatterplot, and fit a regression line to the data.
b Describe the correlation between month and employees.
c What are the limitations of your regression line?

SUMMARY

Scatterplots

- When looking for a relationship between two variables, data can be represented on a scatterplot.
- One variable (the independent variable) is on the x-axis and the other variable (the dependent variable) is on the y-axis.
- Points are plotted by the coordinates formed by each piece of data.
- If the dependent variable consistently increases or decreases as the independent variable increases, a relationship exists.
- If all points on the scatterplot form a straight line, the relationship is said to be linear.
- The pattern of the scatterplot gives an indication of the strength of the relationship or level of association between the variables. This level of association is called correlation.
- A strong correlation between variables does not imply that one variable causes the other to occur.

Median regression lines

- A regression line is the line of best fit on a scatterplot.
- By measuring the gradient and the y-intercept on the regression line, we can use the formula $y = mx + b$ to find the equation.
- When the equation of a regression line has been found, it can then be used to make predictions about the data.
- We can find the regression line by using the eye method or the method of 3-medians.

Correlation

- Correlation is the measure of the relationship between two variables.
- A correlation can be positive or negative and has the same sign as the gradient of the median regression line.
- A positive correlation means that one quantity will increase as the other increases.
- A negative correlation means that one quantity will decrease as the other increases.
- Correlation can be quantified by using a correlation coefficient.
- The correlation coefficient may be interpreted as follows:

$q = 1$	Perfect positive correlation
$0.75 \leq q < 1$	Strong positive correlation
$0.5 \leq q < 0.75$	Moderate positive correlation
$0.25 \leq q < 0.5$	Weak positive correlation
$-0.25 < q < 0.25$	No correlation
$-0.5 < q \leq -0.25$	Weak negative correlation
$-0.75 < q \leq -0.5$	Moderate negative correlation
$-1 < q \leq -0.75$	Strong negative correlation
$q = -1$	Perfect negatve correlation

- The correlation coefficient will always be a number between -1 and 1 or equal to -1 or 1.

CHAPTER REVIEW

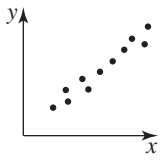
MULTIPLE CHOICE

1 A researcher administers different amounts of fertiliser to a number of trial plots of potato crop. She then measures the total mass of potatoes harvested from each plot. When drawing the scatterplot, the researcher should graph:

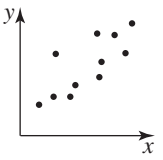
- **A** mass of harvest on the x-axis because it is the independent variable, and amount of fertiliser on the y-axis because it is the dependent variable
- **B** mass of harvest on the y-axis because it is the independent variable, and amount of fertiliser on the x-axis because it is the dependent variable
- **C** mass of harvest on the x-axis because it is the dependent variable, and amount of fertiliser on the y-axis because it is the independent variable
- **D** mass of harvest on the y-axis because it is the dependent variable, and amount of fertiliser on the x-axis because it is the independent variable.

2 Which of the following graphs best depicts a strong negative correlation between variables?

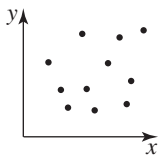
A



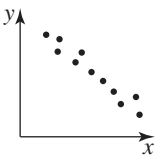
B



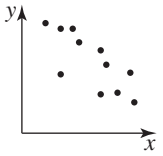
C



D



3 What type of correlation is shown by the graph on the right?



- **A** Strong positive correlation
- **B** Moderate positive correlation
- **C** Moderate negative correlation
- **D** Strong negative correlation

4 A researcher finds that there is a correlation coefficient of –0.62 between the number of pedestrian crossings in a town and the number of pedestrian accidents. The researcher can conclude that:

- **A** Pedestrian crossings cause pedestrian accidents.
- **B** Pedestrian crossings save lives.
- **C** There is evidence to show that pedestrian crossings cause accidents.
- **D** There is evidence to show that the greater the number of pedestrian crossings, the smaller the number of pedestrian accidents.

5 A researcher, who counts the amount of time taken for production line workers to assemble components, relates it to the number of weeks that each worker has spent on the production line. He finds a correlation of –0.82 and can conclude that:

- **A** the greater the number of weeks spent on the production line, the quicker the assembly of components
- **B** the greater the number of weeks spent on the production line, the slower the assembly of components
- **C** many weeks doing the same task causes production workers to become efficient
- **D** many weeks doing the same task causes production workers to become bored and slow as a result.

SHORT ANSWER

1 The table below shows the maximum and minimum temperature on 10 days chosen at random throughout the year. Display this information on a scatterplot.

Maximum temperature (°C)	Minimum temperature (°C)
25	12
36	21
21	11
40	23
24	12
26	15
30	19
18	10
20	8
25	13

2 The table below shows the number of sick days taken by ten employees and relates this to the number of children that they have.

No. of children	No. of sick days
1	5
0	3
3	10
2	8
2	4
4	12
6	12
0	0
1	1
2	2

a Show this information on a scatterplot.
b Does a relationship appear to exist between the number of sick days taken and the number of children they have? If so, is the relationship linear?

3 The table below shows the number of cars and number of televisions in each household.

No. of cars	No. of televisions
1	2
1	1
2	1
2	2
2	0
3	1
1	4
0	3
1	1
2	1

a Show this information on a scatterplot.
b Does a relationship appear to exist between the number of televisions in each household and the number of cars they have? If so, is the relationship linear?

4 The table below shows the relationship between two variables, x and y.

x	2	4	18	7	9	12	2	7	11	10	16
y	103	75	20	66	70	50	95	40	27	42	30

a Prepare a scatterplot of the data.
b On the scatterplot, fit a regression line by eye.
c By measuring the gradient and the y-intercept of the median regression line, find its approximate equation.

5 A survey is conducted comparing household income, I, with house value, V. A scatterplot is drawn and the regression line is found to have the equation $V = 3.7I + 50\,000$. Use the equation to find:
a the likely value of a house owned by a family with an income of $52 000
b the likely income (to the nearest $1000) of a family living in a house valued at $320 000.

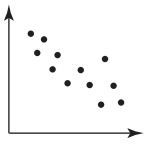
6 An entomologist conducted an experiment in which small amounts of insecticide were introduced to a container of 100 blowflies. The results are detailed below.

Insecticide (I) (micrograms)	No. remaining after 2 h (F)
1	99
2	92
3	81
4	74
5	62
6	68
7	52
8	45
9	38
10	24

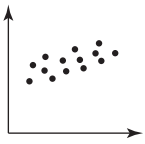
a Display the above information on a scatterplot and, on the scatterplot, draw the median line of regression.
b Find the equation of the regression line.
c Use the equation to predict the number of blowflies that would remain after two hours if 4.25 micrograms of insecticide was introduced.
d Estimate the amount of insecticide needed to remove all blowflies.

7 For each of the following scatterplots, state whether the correlation is positive or negative.

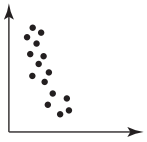
a



b



c



8 The table below shows the relationship between the crowd at cricket matches and the number of matches the home team has won during the season.

No. of wins by home team	Crowd
3	8000
10	21 000
7	11 000
14	22 000
8	13 000
9	12 000
12	19 000

a Display this information on a scatterplot.
b On the scatterplot, draw the median regression line.
c State if a positive or negative correlation exists between the number of wins by the home team and the crowd at their matches.

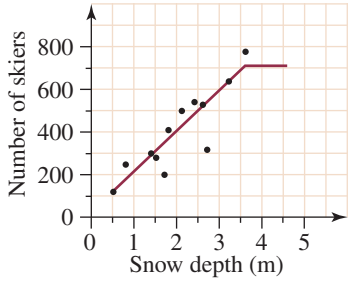
9 For each of the following, state the type of correlation if the correlation coefficient is:
a 0 **b** 1 **c** –0.5
d –0.84 **e** 0.3.

10 An experiment that tested the strength of wooden beams of different thickness demonstrated a correlation of 0.9 between the variables.
a What type of correlation exists in this case?
b What can be said about causality in this case?

11 A survey in which people were asked to state their age and the age of their car revealed a correlation coefficient of –0.65.
a What type of correlation exists in this case?
b What can be said about causality in this case?

EXTENDED RESPONSE

1 The scatterplot on the right shows the number of skiers at a resort and the depth of snow. The median regression line has been drawn on the scatterplot and has the equation $N = 191s + 25$, where N is the number of skiers and s is the snow depth.



a Does a linear relationship exist between depth of snow and number of skiers? Explain your answer.
b Use the equation of the median regression line to estimate:
i the number of skiers if the depth of snow is 3.6 m
ii the depth of snow if there are 500 skiers (correct to 1 decimal place).
c By studying the scatterplot:
i state if the correlation between depth of snow and number of skiers is positive or negative
ii describe the correlation as strong, moderate or weak.

2 The table below shows world population from 1955 to 2005.

Year	1955	1960	1965	1970	1975	1980	1985	1990	1995	2000	2005
World pop. (million)	2750	3000	3400	3700	4000	4400	4800	5300	5750	6073	6451

a Display this information on a scatterplot.
b Draw the median regression line on your scatterplot.
c Let Y be the number of years since 1950 and P the world population. Find the equation of your median regression line.
d Use your graph to estimate the world population in 2010.
e Estimate from your graph when world population will exceed 10 billion.

eBookplus ACTIVITIES

Are you ready?

Digital docs *(page 364)*
- SkillSHEET 12.1 (doc-1421): Finding the median.
- SkillSHEET 12.2 (doc-1422): Using the regression equation to make predictions.
- SkillSHEET 12.3 (doc-1423): Finding the gradient I.
- SkillSHEET 12.4 (doc-1424): Finding the gradient II.

12A Scatterplots

Digital docs
- Spreadsheet (doc-1417): Scatterplot. *(page 368)*
- Spreadsheet (doc-1418): Two variable statistics. *(page 368)*
- Investigation (doc-1419): Collecting bivariate data. *(page 372)*

12B Fitting a straight line by eye

Tutorial
- WE4 int-2441: Fit a straight line to a set of data. *(page 373)*

Digital docs
- WorkSHEET 12.1 (doc-1420): Apply your knowledge of fitting a straight line by eye. *(page 374)*

12C Fitting a straight line — the 3-median method

Tutorial
- WE5 int-2442: Fit a straight line to a set of data. *(page 377)*

Digital docs
- SkillSHEET 12.1 (doc-1421): Finding the median. *(page 381)*
- SkillSHEET 12.2 (doc-1422): Using the regression equation to make predictions. *(page 381)*
- SkillSHEET 12.3 (doc-1423): Finding the gradient I. *(page 381)*
- SkillSHEET 12.4 (doc-1424): Finding the gradient II. *(page 381)*
- Spreadsheet (doc-1425): 3-median regression. *(page 381)*
- Spreadsheet (doc-1426): Making predictions. *(page 381)*
- Investigation (doc-1427): Relationship between variables. *(page 384)*

12D Correlation

Digital docs
- WorkSHEET 12.2 (doc-1428): Apply your knowledge of correlation. *(page 392)*

Chapter review

- Test Yourself (doc-1429): Take the end-of-chapter test to test your progress. *(page 399)*

To access eBookPLUS activities, log on to

www.jacplus.com.au